

Shedding Light on the Asymmetric Learning Capability of AdaBoost

Iago Landesa-Vázquez^a, José Luis Alba-Castro^a

^a*Signal Theory and Communications Department, University of Vigo, Maxwell Street, 36310, Vigo, Spain*

Abstract

In this paper, we propose a different insight to analyze AdaBoost. This analysis reveals that, beyond some preconceptions, AdaBoost can be directly used as an asymmetric learning algorithm, preserving all its theoretical properties. A novel class-conditional description of AdaBoost, which models the actual asymmetric behavior of the algorithm, is presented.

Keywords: AdaBoost, Asymmetry, Boosting, Classification, Cost

1. Introduction

Asymmetry is present in many real world pattern recognition applications. Medical diagnosis, disaster prediction, biometrics, fraud detection, etc. have obviously different costs associated with the different kinds of mistakes (false positives and false negatives) implicitly related to each decision. But asymmetry is not only connected to the direct cost of a mistake. Many problems have unbalanced class priors, where one of the classes is extremely more frequent than the other one, or it is easier to sample. This kind of problems may require classifiers capable of focusing their attention in the rare (but most valuable) class, instead of trying to find hypothesis that in general fit well to data (mainly driven by the prevalent class).

From its original publication, boosting algorithms (Schapire, 1990) and specifically AdaBoost (Freund and Schapire, 1997) have drawn a lot of attention of the pattern recognition community. Its strong properties and theo-

Email addresses: iagolv@gts.uvigo.es (Iago Landesa-Vázquez), jalba@gts.uvigo.es (José Luis Alba-Castro)

retical guarantees, its tendency to non-overfitting and its promising practical results, have focused the interest in this family of algorithms (e.g., Schapire et al., 1998a; Schapire and Singer, 1999; Friedman et al., 2000; Mease and Wyner, 2008a; Viola and Jones, 2004) both from the theoretical (different interpretations, modifications, discussions...) and practical points of view.

In the literature, several modifications of AdaBoost have been proposed to deal with asymmetry (Karakoulas and Shawe-Taylor, 1998; Fan et al., 1999; Ting, 2000; Viola and Jones, 2004, 2002; Sun et al., 2007; Masnadi-Shirazi and Vasconcelos, 2007). Viola and Jones in their face detector framework (2004), use a validation set to modify the AdaBoost strong classifier threshold in order to trade off false positive and detection rates. Nevertheless, as they stated, it is not clear whether this change preserves the original training and generalization guarantees of AdaBoost (Viola and Jones, 2004) and the weak classifiers selection is not optimal for an asymmetric task (Viola and Jones, 2002). Most of the other proposed algorithms (Karakoulas and Shawe-Taylor, 1998; Fan et al., 1999; Ting, 2000; Viola and Jones, 2002; Sun et al., 2007) try to reach asymmetry based on direct manipulations of the weight distribution update rule. These are heuristic modifications of the algorithm, but not a full reformulation of AdaBoost for asymmetric classification problems. On the other hand, the more recent Asymmetric Boosting algorithm (Masnadi-Shirazi and Vasconcelos, 2007) finds a new solution to the problem based on the Statistical View of Boosting (Friedman et al., 2000). Their result is theoretically solid, but the final algorithm is far more complex and computing demanding than the original AdaBoost.

Eventhough some studies (Freund and Schapire, 1997; Zadrozny et al., 2003) mention that the incorporation of unbalanced initial weights could lead to a cost-sensitive version of AdaBoost, subsequent works insist that this is not enough to reach effective asymmetry (Viola and Jones, 2002; Mease et al., 2007; Sun et al., 2007; Masnadi-Shirazi and Vasconcelos, 2007) swelling the number of different asymmetric boosting algorithm variants. Meanwhile, standard AdaBoost remains being explained with an uniform initial weight distribution (e.g., Schapire and Singer, 1999; Friedman et al., 2000; Schapire et al., 1998a; Fan et al., 1999; Ting, 2000; Sun et al., 2007; Masnadi-Shirazi and Vasconcelos, 2007; Freund and Schapire, 1999; Polikar, 2006, 2007). To the best of our knowledge, a formal explanation of the consequences of using asymmetric initial weights on AdaBoost has not been provided, either in one way (they lead to effective asymmetry) or the other (they are definitely useless), so we think that some light must be shed in order to clarify the

actual asymmetric learning capabilities of AdaBoost.

In this paper we propose a new perspective to analyze AdaBoost in a class-conditional way. This analysis suggests that, only with an unbalanced class-conditional initialization of the weight distribution, AdaBoost is, by itself, a theoretically sound asymmetric classification algorithm. Based on class error decomposition, our analysis offers a new model to understand AdaBoost behavior and how it really deals with asymmetry in an additive round-by-round scheme. In fact, weights initialization is no more than a way to modify the data distribution seen by the learner and, as we will see, it can be easily shown to shape the error bound that sets AdaBoost minimization goal. One key point of our work is that it is merely an analysis, so AdaBoost is unchanged. As a consequence, all the algorithm theoretical properties (related to training and generalization errors) remain intact, which has not been clearly reported on the other modifications in the literature. Our analysis is inspired by the generalized derivation of Schapire and Singer (1999), close to the original (Freund and Schapire, 1997) and specially intuitive and illustrative for our purpose. The Statistical View of Boosting (Friedman et al., 2000) and all its subsequent controversy (Mease and Wyner, 2008a; Bennett et al., 2008; Mease and Wyner, 2008b) is left aside, although an analogous conclusion could be derived from it.

The paper is organized as follows: in the next section we will describe AdaBoost original algorithm and its relationship with asymmetry. Section 3 will detail our novel class-dependant interpretation, its analysis and some experimental results which show the actual asymmetric behavior of AdaBoost. Finally, Section 4 includes the main conclusions drawn from this analysis.

2. AdaBoost

In this section we will analyze the original AdaBoost definition and how it has usually been adapted to asymmetric learning.

2.1. Algorithm

Given a set of n training examples (x_i, y_i) from which the m first are positives $\{y_i = 1\}_{i=1}^m$ and the rest are negatives $\{y_i = -1\}_{i=m+1}^n$, AdaBoost is a boosting algorithm whose goal is learning a strong classifier $H(x)$ based on an ensemble of weak classifiers $h_t(x)$ combined in a weighted voting scheme.

$$H(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right) \quad (1)$$

To achieve this, a weight distribution $D_t(i)$ is defined over the whole training set. In each learning round t the weak learner selects the best classifier according to the weight distribution, and this weak classifier is added to the ensemble weighted by a goodness parameter α_t (2) depending on a correlation term r_t (3). Once every weak classifier is selected, the weight distribution is updated according to its performance, following the rule on (4) (where Z_t is a normalization factor which ensures $D_t(i)$ is an actual distribution). The process can be repeated iteratively until a fixed number of rounds is reached, or when the obtained strong classifier achieves some performance goal.

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + r_t}{1 - r_t} \right) \quad (2)$$

$$r_t = \sum_{i=1}^n D_t(i) y_i h_t(x_i) \quad (3)$$

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (4)$$

$$Z_t = \sum_{i=1}^n D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \quad (5)$$

This framework can be seen (Schapire and Singer, 1999) as an additive (round-by-round) minimization process of an exponential bound on the training error of the strong classifier. The bounding process is based on (6), and all the above expressions (including the weight update rule) can be derived from it.

$$H(x_i) \neq y_i \Rightarrow y_i f(x_i) \leq 0 \Rightarrow \exp(-y_i f(x_i)) \geq 1 \quad (6)$$

Following the procedure used by Schapire and Singer (1999), the final bound of the training error obtained by AdaBoost is expressed as (7). The additive minimization of \tilde{E}_T can be seen as finding, round by round, the weak hypothesis h_t that maximizes r_t , that is maximizing the correlation between labels (y_i) and predictions (h_t) weighted by $D_t(i)$.

$$E_T \leq \prod_{t=1}^T Z_t \leq \prod_{t=1}^T \sqrt{1 - r_t^2} = \tilde{E}_T \quad (7)$$

For the sake of simplicity and clarity in our analysis, we will focus on the discrete version of the algorithm. In that case weak hypothesis are binary $y_i \in \{-1, +1\}$, and the minimization process is equivalent to selecting the weak classifier with less weighted error ϵ_t (8)¹. In this case, the last inequality on (7) becomes an equality, and parameter α_t can be rewritten (9) in terms of ϵ_t .

$$\epsilon_t = \sum_{i=1}^n D_t(i) \llbracket h(x_i) \neq y_i \rrbracket = \sum_{\text{nok}} D_t(i) \quad (8)$$

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \quad (9)$$

This simplification doesn't prevent our analysis from being extended to other AdaBoost variations.

2.2. AdaBoost and Asymmetry

AdaBoost is usually seen as a learning procedure driven by misclassification on the training set. In that sense, the exponential bound to minimize must be defined (10) following the guidelines proposed by (Schapire and Singer, 1999). Graphically, we can visualize this bounding process in Figure 1.

$$\begin{aligned} E_T &= \frac{1}{n} \sum_{i=1}^n \llbracket H(x_i) \neq y_i \rrbracket \\ &\leq \frac{1}{n} \sum_{i=1}^n \exp(-y_i f(x_i)) = \prod_{t=1}^T Z_t = \tilde{E}_T \end{aligned} \quad (10)$$

From this point of view, AdaBoost is an algorithm with a symmetric behavior if the number of instances in the training set is the same for the two

¹*Notation:* Operator $\llbracket a \rrbracket$ is 1 if a is true and 0 otherwise. The term 'ok' refers to those training examples in which the result of the weak classifier is right $\{i : h(x_i) = y_i\}$ and 'nok' when it is wrong $\{i : h(x_i) \neq y_i\}$.

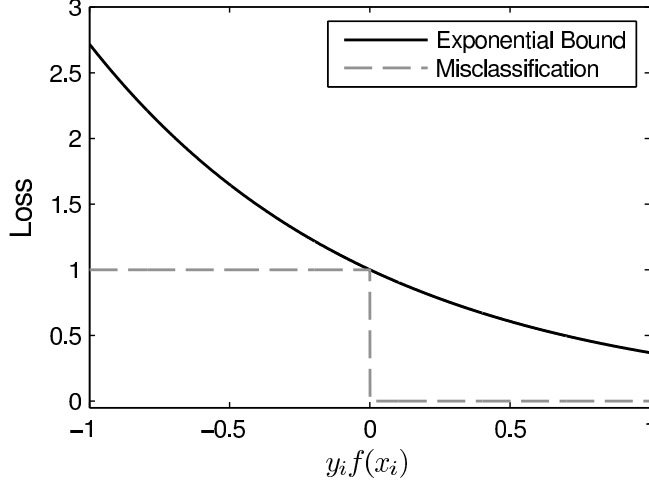


Figure 1: AdaBoost exponential training error bound. Horizontal axis represents the absolute value of the final score of the strong classifier, with negative sign for errors and positive for correct classifications. Vertical axis is the loss related to misclassification and its exponential bound.

classes, or biased to the prevalent class otherwise. Consequently, AdaBoost couldn't be a cost-sensitive algorithm unless the training data set is resampled accordingly.

Based on this seemingly balanced nature, several modifications of AdaBoost have been proposed in order to adapt the algorithm to cost-sensitive problems. Most of them (Karakoulas and Shawe-Taylor, 1998; Fan et al., 1999; Ting, 2000; Sun et al., 2007; Viola and Jones, 2002) are based on modifying the weight update rule in an asymmetric (class-conditional) way. However it is not clear how these changes can affect the theoretical properties of AdaBoost since, as was mentioned above, the update rule is a consequence of the minimization process and not an arbitrary starting point of it.

This perspective is supported by the fact that AdaBoost is usually explained with a fixed uniform initial weight distribution ($D_1(i) = 1/n$) (e.g., Schapire and Singer, 1999; Friedman et al., 2000; Schapire et al., 1998a; Fan et al., 1999; Ting, 2000; Sun et al., 2007; Masnadi-Shirazi and Vasconcelos, 2007; Freund and Schapire, 1999; Polikar, 2006, 2007). Nevertheless some initial works by Freund and Schapire (1997) leave this distribution free to be controlled by the learner. In our explanation of the algorithm in Section 2.1 we deliberately didn't mention anything about the initialization of the weight

distribution. So, what would really happen if the initial distribution was a generic one? Changes of the initial distribution were used to deal with cost-related utility functions (Schapire et al., 1998b), and cost-sensitive weight initializations bonded to different changes in the weight update rule were also used by Karakoulas and Shawe-Taylor (1998), Fan et al. (1999) or Ting (2000). Viola and Jones (2002) proposed a first modification of AdaBoost equivalent to an asymmetric modification of the initial weights. Nevertheless, they discard this approximation arguing that the induced asymmetry is fully absorbed by the first round, remaining the rest of the process entirely symmetric. Their final proposal (coined as *AsymBoost*) was fairly spreading the desired asymmetry among a predefined number of rounds.

Though it is not widely appreciated, it can be easily shown that the error bounded and minimized by AdaBoost is actually a weighted error depending on the initial weight distribution. The only change with regard to the usual bound (10), in which initial uniform weights have been taken out of the summation, is that generic initial weights must be kept inside the summation during the bounding process (11).

$$\begin{aligned}
E_T &= \sum_{i=1}^n D_1(i) \mathbb{I}[H(x_i) \neq y_i] \\
&\leq \sum_{i=1}^n D_1(i) \exp(-y_i f(x_i)) = \prod_{t=1}^T Z_t = \tilde{E}_T
\end{aligned} \tag{11}$$

All the rest of the process remains identical to that explained by Schapire and Singer (1999), consequently guaranteeing all the theoretical properties of AdaBoost with regard to training and generalization errors.

3. Revisiting AdaBoost

In this section we will show our novel class-conditional interpretation model for AdaBoost. This generalized analysis will shed light on the class-dependant behavior of AdaBoost sketched in the previous section.

3.1. Asymmetric Interpretation

To derive our new interpretation of AdaBoost, instead of the initial weight distribution used in the original AdaBoost formulation, we define a set of

parameters which contain exactly the same information of the former distribution.

- *Asymmetry:*

$$\sum_{i=1}^m D_1(i) = \gamma \in (0, 1) \quad (12)$$

- *Class-conditional distributions:*

$$D_{1+}(i) = \frac{D_1(i)}{\gamma}, \quad \text{for } i = 1, \dots, m \quad (13)$$

$$D_{1-}(i) = \frac{D_1(i)}{1 - \gamma}, \quad \text{for } i = m + 1, \dots, n \quad (14)$$

If we put this new set of parameters into the training error expression (11) we will be able to decompose it in terms of its positive and negative class error components:

$$\begin{aligned} E_T &= \sum_{i=1}^n D_1(i) \mathbb{I}[H(x_i) \neq y_i] = \gamma \sum_{i=1}^m D_{1+}(i) \mathbb{I}[H(x_i) \neq y_i] \\ &\quad + (1 - \gamma) \sum_{i=m+1}^n D_{1-}(i) \mathbb{I}[H(x_i) \neq y_i] \\ &= \gamma E_{T+} + (1 - \gamma) E_{T-} \end{aligned} \quad (15)$$

Bounding (15) with the usual exponential approximation, we can also obtain the error bound as the combination of two class-conditional partial error bounds:

$$\begin{aligned} E_T &= \gamma E_{T+} + (1 - \gamma) E_{T-} \\ &\leq \gamma \sum_{i=1}^m D_{1+}(i) \exp(-y_i f(x_i)) \\ &\quad + (1 - \gamma) \sum_{i=m+1}^n D_{1-}(i) \exp(-y_i f(x_i)) \\ &= \gamma \tilde{E}_{T+} + (1 - \gamma) \tilde{E}_{T-} = \tilde{E}_T \end{aligned} \quad (16)$$

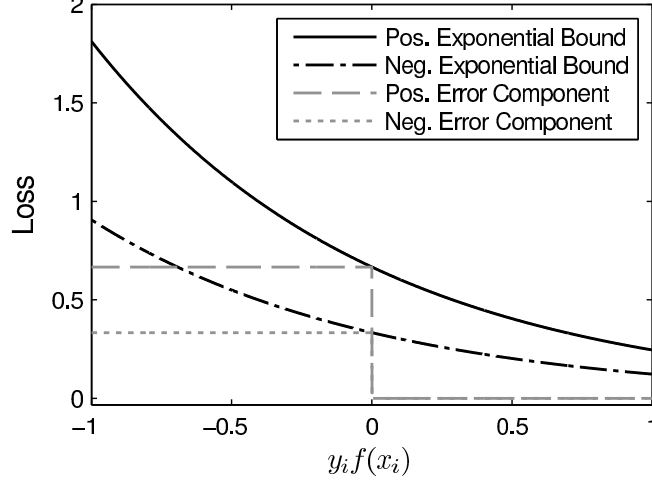


Figure 2: AdaBoost training error and its exponential bound split into two class-conditional components for an asymmetry of $\gamma = 2/3$.

In Figure 2 we can see the defined weighted partial error bounds (\tilde{E}_{T+} and \tilde{E}_{T-}) for an asymmetry of $\gamma = 2/3$ (assuming uniform class-conditional distributions). Asymmetry becomes evident.

As it can be seen, the two partial bounds have expressions formally identical to that of the general bound used in the original AdaBoost (11), so an equivalent update rule can be derived for each class error:

$$D_{(t+1)+}(i) = \frac{D_{t+}(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_{t+}} \quad (17)$$

$$D_{(t+1)-}(i) = \frac{D_{t-}(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_{t-}} \quad (18)$$

where

$$Z_{t+} = \sum_{i=1}^m D_{t+}(i) \exp(-\alpha_t y_i h_t(x_i)) \quad (19)$$

$$Z_{t-} = \sum_{i=m+1}^n D_{t-}(i) \exp(-\alpha_t y_i h_t(x_i)) \quad (20)$$

We will also define two new parameters P_{t+} and P_{t-} which, unraveling the update rules, can be expressed as follows:

$$P_{t+} = \prod_{k=1}^{t-1} Z_{k+} = \sum_{i=1}^m \left(D_{1+}(i) \prod_{k=1}^{t-1} \exp(-\alpha_k y_i h_k(x_i)) \right) \quad (21)$$

$$P_{t-} = \prod_{k=1}^{t-1} Z_{k-} = \sum_{i=m+1}^n \left(D_{1-}(i) \prod_{k=1}^{t-1} \exp(-\alpha_k y_i h_k(x_i)) \right) \quad (22)$$

These parameters (we will discuss later about their meaning) allow us to express the partial error bounds in a compact form:

$$\tilde{E}_{t+} = P_{t+} Z_{t+} \quad (23)$$

$$\tilde{E}_{t-} = P_{t-} Z_{t-} \quad (24)$$

The global error bound of the original view of AdaBoost, \tilde{E}_t , can also be analogously rewritten by defining an equivalent parameter P_t for the whole training set:

$$P_t = \prod_{k=1}^{t-1} Z_k = \sum_{i=1}^n \left(D_1(i) \prod_{k=1}^{t-1} \exp(-\alpha_k y_i h_k(x_i)) \right) \quad (25)$$

$$\tilde{E}_t = P_t Z_t \quad (26)$$

As a result, the error bound to minimize can be expressed as:

$$\begin{aligned} \tilde{E}_t &= \gamma \tilde{E}_{t+} + (1 - \gamma) \tilde{E}_{t-} \\ &= \gamma P_{t+} Z_{t+} + (1 - \gamma) P_{t-} Z_{t-} \end{aligned} \quad (27)$$

Bearing in mind that in each round the only variable parameters are Z_{t+} and Z_{t-} (γ is fixed from the beginning, and P_t depends only on the previous rounds), we can minimize \tilde{E}_t using a procedure analogous to that proposed by Schapire and Singer (1999). While the minimization is exactly the same as in the original case ($\partial \tilde{E}_t / \partial \alpha_t = 0$) the process can be entirely performed in terms of the class-conditional parameters and allows us to obtain the next expression of the error to be minimized round by round:

$$\epsilon_t = \frac{\gamma P_{t+}}{\gamma P_{t+} + (1 - \gamma) P_{t-}} \epsilon_{t+} + \frac{(1 - \gamma) P_{t-}}{\gamma P_{t+} + (1 - \gamma) P_{t-}} \epsilon_{t-} \quad (28)$$

Where ϵ_{t+} and ϵ_{t-} are the partial weighted errors per class:

$$\epsilon_{t+} = \sum_{i=1}^m D_{t+}(i) \llbracket h(x_i) \neq y_i \rrbracket = \sum_{\text{pos nok}} D_{t+}(i) \quad (29)$$

$$\epsilon_{t-} = \sum_{i=m+1}^n D_{t-}(i) \llbracket h(x_i) \neq y_i \rrbracket = \sum_{\text{neg nok}} D_{t-}(i) \quad (30)$$

The expression for α_t is:

$$\begin{aligned} \alpha_t &= \frac{1}{2} \\ &\ln \left(\frac{\gamma P_{t+} \sum_{\text{pos ok}} D_{t+}(i) + (1 - \gamma) P_{t-} \sum_{\text{neg ok}} D_{t-}(i)}{\gamma P_{t+} \sum_{\text{pos nok}} D_{t+}(i) + (1 - \gamma) P_{t-} \sum_{\text{neg nok}} D_{t-}(i)} \right) \\ &= \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \end{aligned} \quad (31)$$

And the final training error bound, can be expressed as:

$$E_T \leq \tilde{E}_T = \prod_{t=1}^T Z_t = \prod_{t=1}^T \sqrt{1 - r_t^2} \quad (32)$$

$$\begin{aligned} r_t &= \frac{\gamma P_{t+}}{\gamma P_{t+} + (1 - \gamma) P_{t-}} \sum_{i=1}^m D_{t+}(i) y_i h_t(x_i) \\ &\quad + \frac{(1 - \gamma) P_{t-}}{\gamma P_{t+} + (1 - \gamma) P_{t-}} \sum_{i=m+1}^n D_{t-}(i) y_i h_t(x_i) \end{aligned} \quad (33)$$

As we can see, all the magnitudes (ϵ_t , α_t and r_t) are systematically decoupled in two components according to the global asymmetry and the classifier behavior over each class. The key concept is that expressions (28), (31) and

(33) are actually the same as those of the original AdaBoost formulation (8), (9) and (3), respectively. On one hand, the derivation is equivalent to the original with the only exception that weights are decomposed in three parameters (12), (13), (14). On the other hand, during the derivation process we can obtain equivalences (34), (35), (36) which appropriately replaced on the original AdaBoost expressions lead us to the new ones.

$$\gamma P_{t+} + (1 - \gamma) P_{t-} = P_t \quad (34)$$

$$\gamma P_{t+} D_{t+}(i) = P_t D_t(i), \text{ for } i = 1, \dots, m \quad (35)$$

$$(1 - \gamma) P_{t-} D_{t-}(i) = P_t D_t(i), \text{ for } i = m + 1, \dots, n \quad (36)$$

3.2. Asymmetric Error Analysis

The initial weight decomposition in our analysis allows us to decouple the global weight distribution information in two levels which were always mixed in the original AdaBoost formulation:

- *Class level:* The asymmetry parameter γ models the global cost of the positive class over the negative one. From a practical point of view, this parameter can be used to introduce asymmetry in the strong classifier.
- *Example level:* The class-conditional initial weight distributions (D_{1+} and $D_{1-}(i)$) model the relative relevance of each example inside its own class. So, being two separate distributions, they are isolated from the asymmetry of the problem.

This two-level categorization can be extrapolated to the error bound minimized by AdaBoost in each round, yielding us a new insight.

$$\begin{aligned}
\tilde{E}_t = & \underbrace{\underbrace{\underbrace{\gamma}_{\text{Global Asymmetry}} \cdot \underbrace{P_{t+}}_{\text{Previous Rounds Asymmetry}}}_{\text{Effective Asymmetry}} \cdot \underbrace{Z_{t+}}_{\text{Current Round}}}_{\text{CLASS LEVEL} \quad \text{EXAMPLE LEVEL}} \\
& \underbrace{\hspace{10em}}_{\text{POSITIVES BEHAVIOR}} \\
& + \underbrace{\underbrace{\underbrace{(1-\gamma)}_{\text{Global Asymmetry}} \cdot \underbrace{P_{t-}}_{\text{Previous Rounds Asymmetry}}}_{\text{Effective Asymmetry}} \cdot \underbrace{Z_{t-}}_{\text{Current Round}}}_{\text{CLASS LEVEL} \quad \text{EXAMPLE LEVEL}} \\
& \underbrace{\hspace{10em}}_{\text{NEGATIVES BEHAVIOR}}
\end{aligned} \tag{37}$$

The bound consists of two formally identical terms, one per class (positive and negative). Each term has two main components: one on the class level and another one in the example level.

- The *class level* defines the effective asymmetry demanded for the current round. It can be seen as the global desired asymmetry modulated by the past asymmetric behavior of the classifier (encoded by cumulative errors P_{t+} and P_{t-}). It only depends on the previous rounds.
- The *example level* is related to the weighted error of the current weak classifier. Weight distributions ($D_{t+}(i)$ and $D_{t-}(i)$) are updated, round by round, to encode the effective relative relevance of each example totally apart from the class behavior. It depends both on the previous and current rounds.

As we can see, the effective asymmetry of each round will depend on the asymmetry of the previous ones, so AdaBoost goal is to iteratively find the weak hypothesis which, given its predecessors, best helps to the global asymmetry minimizing the training error bound. Asymmetry is reached in a round-by-round adaptive way, without any previous restriction on the final number of rounds.

This error bound interpretation can open the door to new modifications of AdaBoost based, for example, on tuning the global and past asymmetry contributions in order to achieve different asymmetric behaviors along rounds.

3.3. Algorithm

Once we have seen the actual asymmetric properties of AdaBoost when using a generic initial distribution, the complete algorithm can be reformatted as in Table 1.

The only change regarding to the algorithm description usually found in the literature is that the initial weight distribution is not necessarily uniform. Here, we initialize it in terms of an asymmetry parameter (γ) and two class-conditional distributions ($D_{1+}(i)$ and $D_{1-}(i)$), which can be uniform (all the examples of each class weight the same) or not (some examples are emphasized).

3.4. Experiments

In order to illustrate our analysis with empirical results on the asymmetric behavior of AdaBoost with unbalanced initial weight distributions, we performed three kinds of experiments. For these experiments we have defined the Asymmetric Error (AsErr) as the cost-sensitive error of the classifier: the weighted average of positives (PosErr) and negatives (NegErr) error rates or, what is the same, the weighted average of false negatives (FN) and false positives (FP) rates.

$$\begin{aligned}\text{AsErr} &= \gamma \cdot \text{PosErr} + (1 - \gamma) \cdot \text{NegErr} \\ &= \gamma \cdot \text{FN} + (1 - \gamma) \cdot \text{FP}\end{aligned}\tag{38}$$

At first, we used the separable set of Figure 3 (inspired by that used by Viola and Jones, 2002) in which positives are concentrated in a circular area and negatives surround them, following the same uniform distribution in both cases. Weak classifiers are stumps in the linear two-dimensional space.

AdaBoost behavior for this training set and different asymmetries ($\gamma = \frac{1}{2}$, $\frac{3}{5}$, $\frac{2}{3}$ and $\frac{7}{8}$) is shown in Figure 4. We can see that, as the asymmetry grows, positive error bound and respective positive training/test errors tend to be lower, while negative error bound and respective negative training/test errors tend to be higher. This behavior doesn't prevent the classifier from asymptotically improving itself round by round approaching to zero training error

Table 1: Discrete AdaBoost generalized formulation for asymmetric classification problems.

Given:

- A set of positive examples: $(x_i, y_i) = (x_1, 1), \dots, (x_m, 1)$
- A set of negative examples: $(x_i, y_i) = (x_{m+1}, -1), \dots, (x_n, -1)$
- An asymmetry parameter: $\gamma \in (0, 1)$.
- Two weight distributions over the positive ($D_{1+}(i)$) and negative examples ($D_{1-}(i)$).

Initialize the global weight distribution as:

- $D_1(i) = \gamma D_{1+}(i)$ for $i = 1, \dots, m$
- $D_1(i) = (1 - \gamma) D_{1-}(i)$ for $i = m + 1, \dots, n$

For $t = 1, \dots, T$ (or until the strong classifier reaches some performance goal):

- Select the weak classifier $h_t(x)$ with the lowest weighted error

$$\epsilon_t = \sum_{i=1}^n D_t(i) \mathbb{I}[h_t(x_i) \neq y_i] = \sum_{\text{nok}} D_t(i)$$

- Calculate

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

- Update the weight distribution

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{\sum_{i=1}^n D_t(i) \exp(-\alpha_t y_i h_t(x_i))}$$

The final strong classifier is:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$$

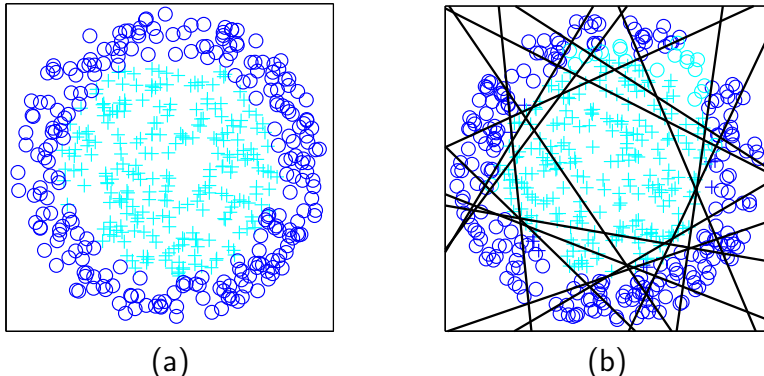


Figure 3: Training set (a) and example weak classifiers over the test set (b) used to illustrate our asymmetric analysis of AdaBoost. Positive examples are marked as ‘+’, while ‘o’ are the negative ones.

classifiers, due to the separable nature of the classification problem. The key advantage of this approach is that the error evolution follows an unbalanced behavior, allowing the user to stop training at any iteration, with the theoretical confidence of having minimized the error bound with the desired asymmetry no matter in which iteration we are (opposed to Asymboost (Viola and Jones, 2002) philosophy). This can be very useful for flexible building of cascaded classifiers as the ones proposed by (Viola and Jones, 2004).

We also run this experiment with a non-separable set as shown in Figure 5 and for the same different asymmetries ($\gamma = \frac{1}{2}, \frac{3}{5}, \frac{2}{3}$ and $\frac{7}{8}$) (Figures 6 and 7). We can see that, due to the overlapping between classes (they are non-separable), error curves tend to a working point different to that of the previous experiment. In any case, the obtained behaviors are clearly asymmetric along the whole evolution of the boosted classifiers, and the degree of asymmetry is effectively managed by the γ parameter.

Finally we have also conducted a more extensive experiment using both synthetic and real datasets to obtain numerical results verifying our hypothesis. The strategy we have followed is *leave-one out cross-validation*. Thus, iteratively selecting every example of a dataset, a classifier is trained over the remaining elements and tested over the selected one. This procedure is repeated for all the examples, all the datasets and all the desired γ parameters, so that overall performance figures can be computed. Tables 2 and 5 summarize the obtained performance over the synthetic dataset with overlapping in Figure 5 and some real asymmetric datasets (Credit, Diabetes

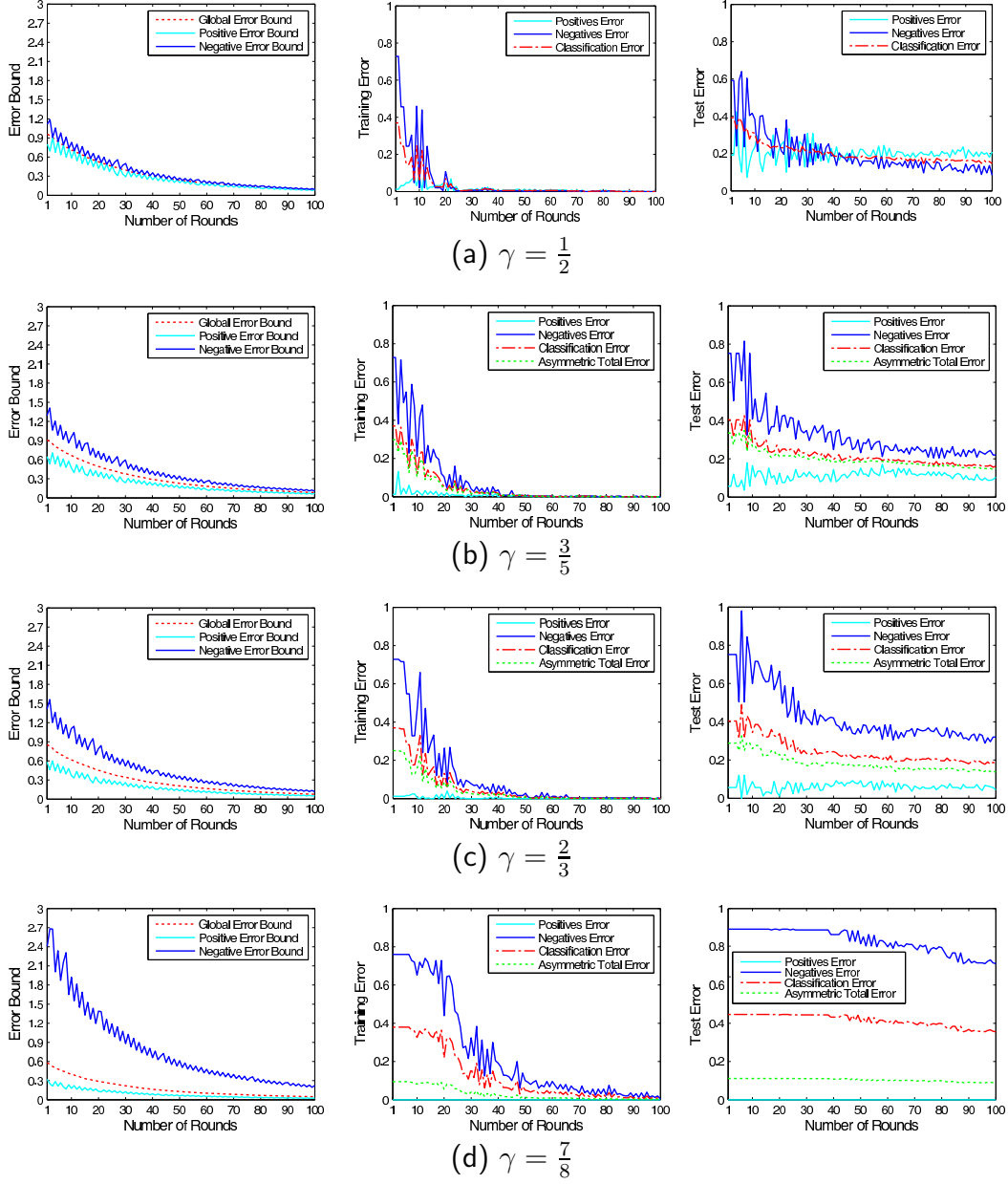


Figure 4: Evolution of training error bounds (left column), training errors (center column) and test errors (right column) through 100 rounds of AdaBoost training and different asymmetries, using the set without overlapping in Figure 3.

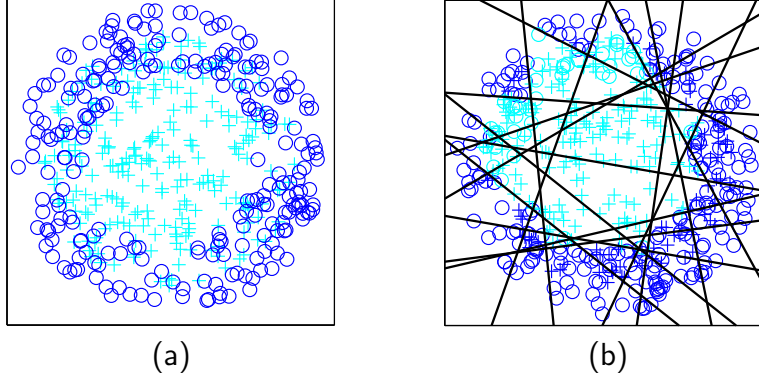


Figure 5: Training set with overlapping (a) and example weak classifiers over the test set (b). Positive examples are marked as '+', while 'o' are the negative ones.

and Spam) extracted from the UCI Machine Learning Repository (Frank and Asuncion, 2010). As can be seen, in all cases a consistent asymmetric behavior is reached, being also progressive depending on γ .

Table 2: Classifier behavior (false negatives, false positives, classification error and asymmetric error) for different asymmetric requirements over the synthetic cloud dataset with overlapping in Figure 5.

γ	Synthetic cloud			
	FN	FP	ClErr	AsErr
1/2	31.60%	29.20%	30.40%	30.40%
3/5	26.80%	38.00%	32.40%	31.28%
2/3	22.00%	42.00%	32.00%	28.67%
7/8	7.60%	66.40%	37.00%	14.95%

Table 3: Classifier behavior (false negatives, false positives, classification error and asymmetric error) for different asymmetric requirements over real datasets extracted from the UCI Machine Learning Repository (Frank and Asuncion, 2010).

γ	Credit				Diabetes				Spam			
	FN	FP	ClErr	AsErr	FN	FP	ClErr	AsErr	FN	FP	ClErr	AsErr
1/2	28.67%	26.86%	27.40%	27.76%	32.09%	22.40%	25.78%	27.24%	4.84%	6.18%	5.37%	5.51%
3/5	22.67%	37.43%	33.00%	28.57%	22.39%	28.60%	26.43%	24.87%	4.16%	7.06%	5.30%	5.32%
2/3	18.67%	43.43%	36.00%	26.92%	19.78%	32.20%	27.86%	23.92%	3.84%	8.38%	5.63%	5.35%
7/8	6.00%	69.14%	50.20%	13.89%	10.07%	53.00%	38.02%	15.44%	2.33%	11.75%	6.04%	3.51%

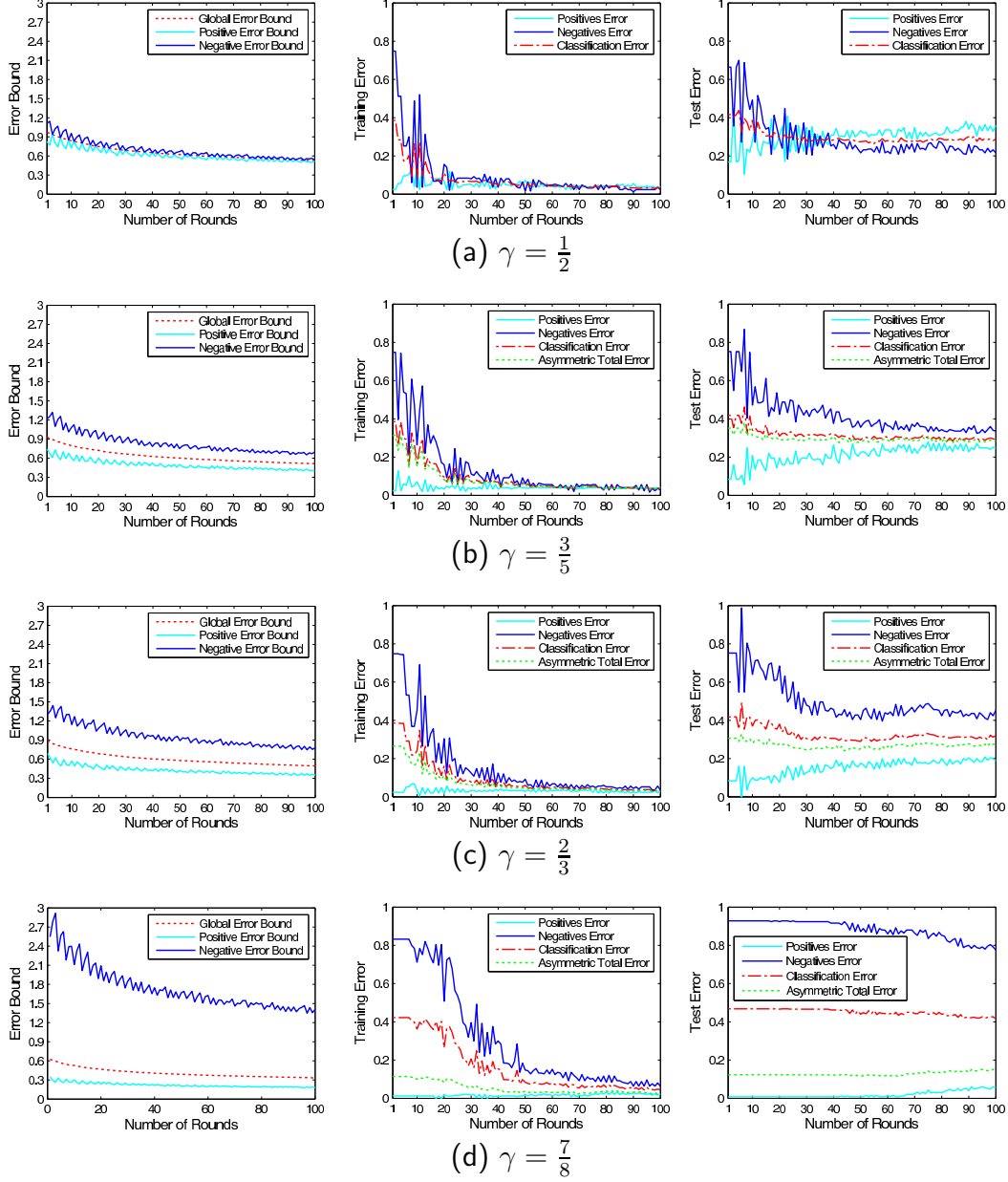


Figure 6: Evolution of training error bounds (left column), training errors (center column) and test errors (right column) through 100 rounds of AdaBoost training and different asymmetries, using the set with overlapping in Figure 5.

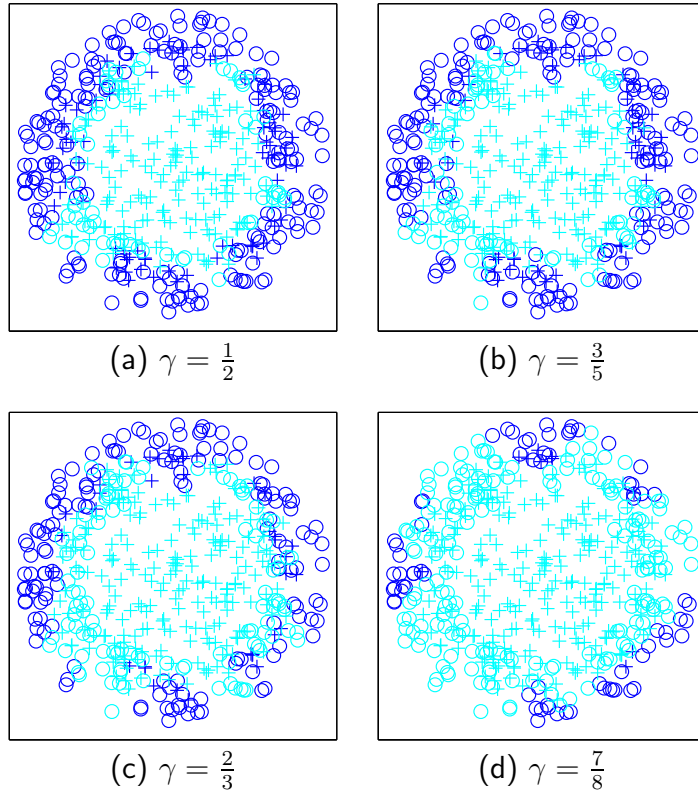


Figure 7: Classification results over the test set with overlapping (Figure 5) for different asymmetries. As in Figure 3 true positives are marked as '+', and 'o' are true negatives. However, in this case, cyan colored marks represent positive classifications while blue ones represent negative classifications.

3.5. Discussion

Previous sections reveal that AdaBoost can be by itself an asymmetric learning algorithm, following its original additive round-by-round updating behavior. Our proposed change of perspective yields several consequences:

- The initial weight distribution is more than the distribution seen by the first weak classifier. It is the distribution which weighs the global error bound to be minimized by AdaBoost. Any asymmetry in this initial weight distribution is an effective way to introduce asymmetry in the strong classifier goal.
- This kind of asymmetry is asymptotic for the whole classifier and the number of training rounds can be as flexible as in the original case (unlike AsymBoost, Viola and Jones, 2002, which rigidly spreads the asymmetry in a predefined number of rounds). Among other advantages, this makes possible, once a strong classifier is trained, to cut it out at whatever round we consider, with the certainty that the error bound has been minimized taking the desired global asymmetry into account. Moreover, it can be specially useful for cascaded classifiers as those used for object detection (Viola and Jones, 2004), in which each stage (each strong classifier) must be markedly asymmetric and as short as possible, in order to improve rejecting efficiency (and consequently the real-time ability of the system).
- Asymmetry can be reached without changing the weight update rule, as opposed to the most of the asymmetric AdaBoost modifications in the literature. It is argued that such a modification is needed because AdaBoost updates weights of examples from different classes in the same way, only distinguishing between correctly and incorrectly classified ones. This is true, but it must be taken into account that, before the first weight distribution update, AdaBoost must have selected a first weak classifier $h_1(x)$ and a goodness parameter α_1 according to the initial weight distribution $D_1(i)$, which stores the desired asymmetry information. Consequently $h_1(x)$ and α_1 implicitly encode asymmetry information, and both parameters are just the ones that manage the update rule. The result is that asymmetry is indirectly present in the usual weight update rule and, as seen in section 3.2, all the subsequent

iterations can be seen as a round-by-round asymmetry adaptive process. Any additional class-dependant change in the weight update rule may emphasize, in a more or less controlled way, the described asymmetric behavior but in those cases it is not clear how it would affect to the theoretical properties of AdaBoost.

- The whole formal guarantees provided by AdaBoost remain intact.

4. Conclusion

In this paper we have introduced a new insight on the asymmetric learning capabilities of AdaBoost, in which the symmetric case can be seen as a particularization (when asymmetry parameter $\gamma = 0.5$). Beyond some preconceptions, the only needed change with regard to the usual formulation is how the initial weights are initialized. We have shown, using a novel class-conditional interpretation of the error bound, that the asymmetric behavior reached is asymptotic with the number of rounds and it works, as the whole algorithm, in an additive round-by-round way. The weight update rule doesn't need to be changed and all the formal guarantees remain intact. Our error bound interpretation can also be useful to develop new AdaBoost modifications based on adjusting the different asymmetry components (both on the class and/or example levels). We have not presented a new algorithm...it is just AdaBoost!

5. Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation through project TEC2008-05894 and by the Galician Government through grants IN840C, IN808C and CN2011/019.

References

Bennett, K., Buja, A., Stuetzle, W., Freund, Y., Schapire, R., Friedman, J., Hastie, T., Tibshirani, R., Bickel, P., Ritov, Y., Bhlmann, P., Yu, B., 2008. Responses to evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research* 9, 131–156.

- Fan, W., Stolfo, S., Zhang, J., Chan, P., 1999. Adacost: Misclassification cost-sensitive boosting, in: Proc. 16th International Conference on Machine Learning, pp. 97–105.
- Frank, A., Asuncion, A., 2010. Uci machine learning repository.
- Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119–139.
- Freund, Y., Schapire, R., 1999. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* 14, 771–780.
- Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* 28, 337–407.
- Karakoulas, G., Shawe-Taylor, J., 1998. Optimizing classifiers for imbalanced training sets, in: *Advances in Neural Information Processing Systems* 12, pp. 253–259.
- Masnadi-Shirazi, H., Vasconcelos, N., 2007. Asymmetric boosting, in: Proc. 24th International Conference on Machine Learning, pp. 609–619.
- Mease, D., Wyner, A., 2008a. Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research* 9, 175–194.
- Mease, D., Wyner, A., 2008b. Evidence contrary to the statistical view of boosting: A rejoinder to responses. *Journal of Machine Learning Research* 9, 195–201.
- Mease, D., Wyner, A., Buja, A., 2007. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research* 8, 409–439.
- Polikar, R., 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 21–45.
- Polikar, R., 2007. Bootstrap-inspired techniques in computational intelligence. *IEEE Signal Processing Magazine* 24, 59–72.
- Schapire, R., 1990. The strength of weak learnability. *Machine Learning* 5, 197–227.

- Schapire, R., Freund, Y., Bartlett, P., Lee, W., 1998a. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* 26, 1651–1686.
- Schapire, R., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning* 37, 297–336.
- Schapire, R., Singer, Y., Singhal, A., 1998b. Boosting and rocchio applied to text filtering, in: *Proc. ACM SIGIR*, pp. 215–223.
- Sun, Y., Kamel, M., Wong, A., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40, 3358–3378.
- Ting, K., 2000. A comparative study of cost-sensitive boosting algorithms, in: *in Proc. 17th International Conference on Machine Learning*, pp. 983–990.
- Viola, P., Jones, M., 2002. Fast and robust classification using asymmetric adaboost and a detector cascade, in: *Advances in Neural Information Processing Systems* 14, pp. 1311–1318.
- Viola, P., Jones, M., 2004. Robust real-time face detection. *International Journal of Computer Vision* 57, 137–154.
- Zadrozny, B., Langford, J., Abe, N., 2003. Cost-sensitive learning by cost-proportionate example weighting, in: *Proc. IEEE International Conference on Data Mining (ICDM)*, pp. 435–442.